Department of Electrical and Electronic Engineering
Khulna University of Engineering & Technology
Khulna-9203, Bangladesh

Course No: EE 3122
Sessional on Numerical Methods & Statistics

Experiment No. 1
Name of the Experiment: Statistical Analysis of Data
Objectives:

[1] To read data (txt / Excel) from a file and write data to another format of file
[2] To convert ungrouped data into grouped data
[3] To calculate the number of data in a column and/or row of a file (ASCII file/Excel file)
[4] To calculate mean (AM, GM, HM), mode, median, from grouped and ungrouped data
[5] To calculate standard deviation and variance from grouped and ungrouped data
[6] To determine the moments of a distribution

Theory/Introduction:

Measures of Central Tendency

- Arithmetic Mean – average, Geometric Mean, Harmonic Mean, Median, Mode

Arithmetic mean: Most important measure of central location is the mean or average value for a set of variable. The arithmetic mean is the sum of a set of observations, positive, negative or zero, divided by the number of observations. If we have "n" real numbers x_1, x_2, x_n, then their arithmetic mean, denoted by $\overline{x}$, can be expressed as:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{(Ungrouped data)}$$

$$\overline{x} = \frac{\sum_{i=1}^{n} f_i x_i}{n} \quad \text{(Grouped data)}$$

x_i =value of i-th item (Ungrouped data)
$\overline{x}$ =arithmetic mean
x_i =mid value of i-th class (Grouped data)
f_i =frequency of i-th class (Grouped data)
n = total frequency

Geometric Mean:

- Geometric mean is defined as the positive root of the product of observations. Symbolically,

$$G = (x_1 x_2 ... x_n)^{\frac{1}{n}} \quad \text{(Ungrouped data)}$$

$$G = anti\log\left(\frac{1}{n}\sum_{i=1}^{n} \log x_i\right) \quad \text{(Ungrouped data)}$$

- If the "n" non-zero and positive variate-values $x_1, x_2, \ldots, x_n$ occur $f_1, f_2, \ldots, f_n$ times, respectively, then the geometric mean of the set of observations is defined by:

$$G = \left(x_1^{f_1} x_2^{f_2} ... x_n^{f_n}\right)^{\frac{1}{n}} = \left[\prod_{i=1}^{n} x_i^{f_i}\right]^{\frac{1}{n}}, \text{where } n = \sum_{i=1}^{n} f_i \quad \text{(Grouped data)}$$

$$G = anti\log\left(\frac{1}{n}\sum_{i=1}^{n} f_i \log x_i\right) = anti\ln\left(\frac{1}{n}\sum_{i=1}^{n} f_i \ln x_i\right) \quad \text{(Grouped data)}$$

Harmonic Mean: The harmonic mean H of the positive real numbers $x_1, x_2, \ldots, x_n$ is defined to be

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$ (Direct method, ungrouped data)

$$H = \frac{n}{\frac{f_1}{x_1} + \frac{f_2}{x_2} + \ldots + \frac{f_n}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{f_i}{x_i}}$$ (Direct method, Grouped data)

Median: The **median** of a set of data values is the middle value of the data set when it has been arranged in ascending order. That is, from the smallest value to the highest value.

If "n" is odd

$$M_e = x_{\frac{1}{2}(n+1)}$$

If "n" is even

$$M_e = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right)$$ (Ungrouped data)

Median of grouped data:

$$Median = L_1 + \frac{\frac{n+1}{2} - f_c}{f_m} \times C$$ (Grouped data)

L_1 = lower limit of median group
C = class interval of median group/ width of the median class
n = total frequency
f_m = frequency of the median group
f_c = cumulative frequency of the group preceding the median group/ Cumulative frequency of the pre-median class

Procedure:
(i) Locate the median class by cumulating class frequency
(ii) Find the class frequency whose cumulative frequency first exceeds or equal to (n+1)/2. This class contains the median and is called median class.
(iii) Estimate the median with the median class.

Mode: Mode of Group Data:

$$Mode = L_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$$ (Grouped data)

- L_1 = Lower boundary of modal class
- Δ_1 = difference of frequency between modal class and pre-modal class
- Δ_2 = difference of frequency between modal class and post-modal class
- C = class interval

Modal class is the class which has highest frequency.

Calculation of Mode: An empirical formula,

Mode=Mean – 3(Mean- Median) = 3Median-2Mean

Standard Deviation: The Standard Deviation is a measure of how spread out numbers is. Its symbol is σ (the Greek letter sigma). The formula is easy: it is the square root of the Variance. So now you ask, "What is the Variance?" The Variance is defined as: The average of the squared differences from the Mean.

- $$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \left\{\frac{\sum_{i=1}^{n} x_i}{n}\right\}^2}$$ (Ungrouped data)

- $$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}{n}}, \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i x_i^2}{n} - \left\{\frac{\sum_{i=1}^{n} f_i x_i}{n}\right\}^2}$$ (Grouped data)

- N.B. Variance = standard deviation^2

Measures of Moments:

- Moments are constant values in a given distribution which help us to ascertain the nature and form of a distribution
- A moment is a quantitative measure of the shape of a set of points.
- The **first moment is called the mean** which describes the center of the distribution.
- The **second moment is the variance** which describes the spread of the observations around the center, 2nd moment about the mean = variance
- **3rd moment - Skewness** (describes asymmetry and the **4th moment - Kurtosis** (describes peakedness)

Formula of moments for Grouped data:

- **rth central moment about mean,** $\mu_r = \frac{\sum f_i(x_i - \bar{x})^r}{n}$, where $n = \sum f$
- **rth moment about any arbitrary value,** $\nu_r = \frac{\sum f_i(x_i - A)^r}{n}$, where $n = \sum f$
- **rth moment about zero value,** $\nu_r = \frac{\sum f_i x_i^r}{n}$, where $n = \sum f$
- Here, x_i=mid value of class, f=frequency of each class, n=total frequency

Formula of moments for ungrouped data:

- ***r*-th central moment about mean,** $\mu_r = \frac{\sum(x_i - \bar{x})^r}{n}$
- ***r*-th moment about any arbitrary value,** $\nu_r = \frac{\sum(x_i - A)^r}{n}$
- ***r*-th moment about zero value,** $\nu_r = \frac{\sum x_i^r}{n}$, where $n = \sum f$
- Here, x_i=individual data and n=total frequency

Experiment:

(1) Perform the mean (AM, GM, HM), median, mode, SD, Variance, moments (about mean and about any arbitrary value) from the following ungrouped data using C/C++ or MatLab.

Sample Input Data: (**indata1.txt or indata1.xls**). Put the data in your input file as a column data or row data with spacing, or comma delimited.

5, 12, 5, 7, 21, 23, 24, 15, 8, 22, 23, 9, 11, 18, 22, 21, 19, 3, 4, 20

(2) Perform the mean (AM, GM, HM), median, mode, SD, Variance, moments (about mean and about any arbitrary value) from the grouped data using C/C++ or MatLab (**Column 1 and 3 are used as input data, use the file number indata2.txt or indata2.xls**).

Marks	No. of students, fi
30-40	3
40-50	7
50-60	20
60-70	25
70-80	20
80-90	15
90-100	10

Here, n=100. Therefore, (n+1)/2 = 50.5. The value of 50.5th item is the median. Median holding lies in the group (60-70) since the cumulative frequency of this group just exceeds 50.5. By using formula, Median = 60 + (50.5-30)/25×10=60+8.2=68.2

Report: Prepare your report according to following fashion.
(Red color, your main part)

Cover page – Your overall information

Expt. No.
Name of Experiments:
Objectives
Introductory information

Calculations, Curve, graphs, Discussions
Conclusions

Reference (if any)